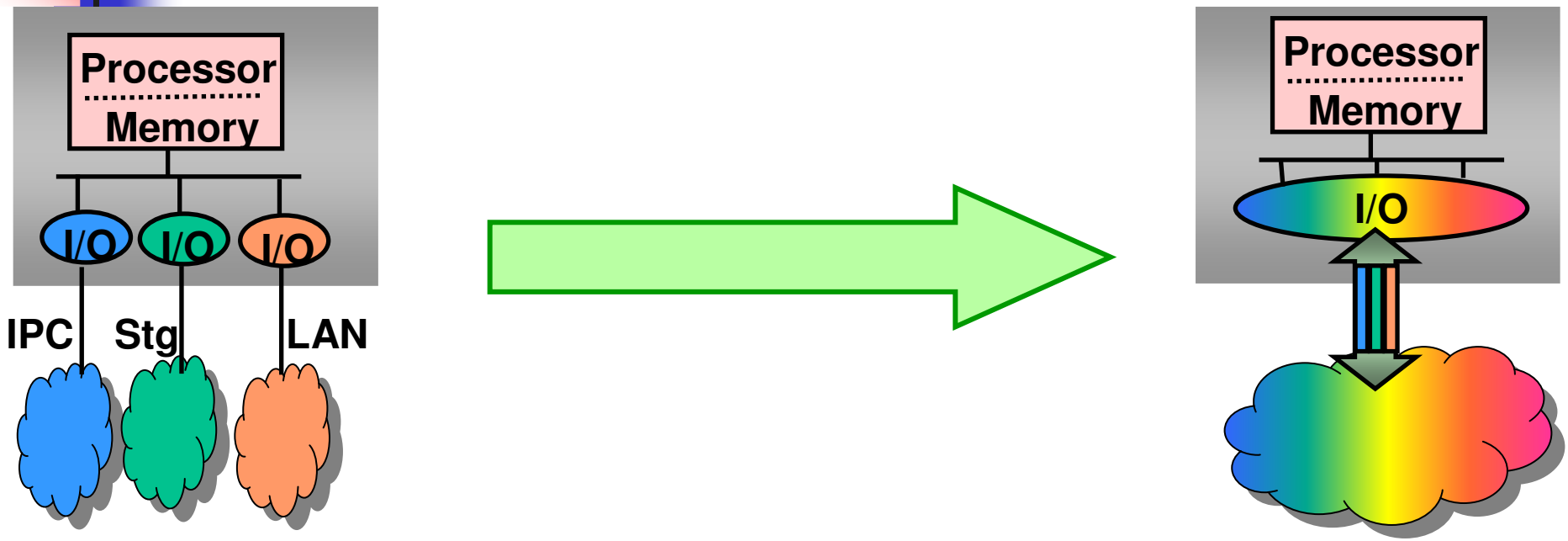
Fabric Convergence from a Storage Perspective

Mike Ko, IBM
Manoj Wadekar, Intel
Davide Bergamasco, Cisco
Joe Pelissier, Brocade

May 29, 2007

I/O Consolidation in the Datacenter



- Enhancing Ethernet to enable I/O consolidation in the datacenter has been discussed in 802 meetings since 2004
- Proposals on congestion management are currently being debated in 802.1Qau working group

Key Enablers for Storage Convergence

- For I/O consolidation onto a single enhanced Ethernet link
  - Storage will be sharing the link with other classes of applications such as IPC and LAN
- There are 3 key areas that need to be considered:
  - Priority processing and packet scheduling
  - Per priority flow control (e.g. PAUSE)
  - Discovery and capability exchange protocol
- This will be crucial for new emerging storage protocols
  - Discussions are underway at T11 for layering Fibre Channel directly over Ethernet

Priority Processing and Packet Scheduling

- For each class of applications that will now use the same consolidated layer 2 transport
  - Queuing requirements for different traffic classes are needed to allow for different resource allocation
- Different traffic classes need to be managed separately
  - LAN
    - Large number of flows, not very sensitive to latency
    - E.g. dominant traffic type in Front End Servers
  - SAN
    - Large packet sizes, sensitive to packet drops
    - E.g. Middle Tier and Back End Servers
  - IPC:
    - Mix of large and small messages
      - Small messages are latency sensitive
    - E.g. Back End Servers, HPC Applications

Use of Queuing Requirements in Storage

- Priority groups allow storage traffic to be managed as a group with configurable QOS guarantees
  - Ensures that storage traffic will get its fair share of resources
  - Allows the scheduling mechanism to apply different disciplines
    - Provide minimal latency for delay sensitive traffic in other bandwidth groups
- If necessary, different queues can be set up within the storage traffic class group with different QOS allocation

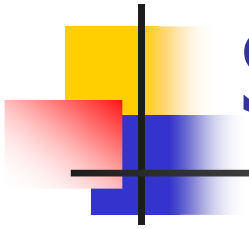
Proposals on Priority Processing and Packet Scheduling

- We need to start developing a list of requirements/objectives for priority processing and packet scheduling
- Proposals made in the following presentations can be used as a basis for the draft:
  - "Improved Transmission Selection" by Wadekar et al, May '06
    - new_cm_wadekar_transmission_selection-0506-01
  - "Proposal for Traffic Differentiation in Ethernet Networks" by Wadekar et al, March '05
    - new-wadekar-virtual-links-0305
  - "Congestion Management in Datacenter Networks" by Wadekar, May '05
    - new-wadekar-congestion-management-framework-0505
  - "Proposal to improve expedited forwarding" by Congdon, May '05
    - new-congdon-improved-queuing-0505
  - "Proposal to improve expedited forwarding (...continued...)" by Congdon, July '05
    - new-congdon-improved-queuing-0705

Per Priority Flow Control and Storage

- Scheduling mechanism allows storage traffic to share the same link with non storage traffic
  - But to achieve the no packet drop behavior required by some storage protocols, per priority flow control (e.g., PAUSE) will be needed
- Per Priority PAUSE extends the granularity of 802.3x PAUSE mechanism to accommodate different priority classes
  - Selective pausing avoids impacts to high priority and delay sensitive traffic
  - For storage protocols layered over TCP/IP, per priority flow control enables service differentiation at the link layer (vs at the IP layer)

Impact of Dropped Packets in Storage



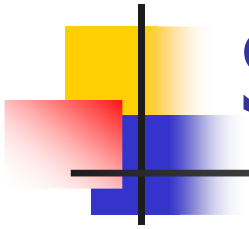
- For storage traffic that uses TCP/IP as the transport, e.g., iSCSI, iSER, etc.
  - Besides retransmission delay, TCP/IP also exhibits additive-increase-multiplicative-decrease (AIMD) behavior in response to packet drops
    - Hurts throughput and latency
  - Alternatives with different congestion avoidance algorithms include FastTCP, HighSpeed TCP, BIC-TCP, H-TCP, XCP, etc.
- For storage traffic that does not use a transport layer, e.g., Fibre Channel over Ethernet
  - Detection at the SCSI level is in the order of 10s of seconds
    - Detection time is in the order of seconds if Read Exchange Concise (REC) extended link service is supported
  - Recovery is at the SCSI command level
    - Severely hurts throughput and latency
    - May cause severe system malfunction (e.g., unexpected server reboots)

Proposals on Per Priority PAUSE

- We need to start developing a list of requirements/objectives for per priority flow control
- Proposals made in the following presentations can be used as a basis for the draft:
  - "Why Priority/Class Based PAUSE is Required" by Brunner et al, July '05
    - brunner_1_0507
  - "Priority Pause support for CN (e.g., BCN) Mechanism" by Hazarika et al, November '06
    - au-Brunner-Hazarika-Priority-Pause-considerations-111406
- Concerns about deadlocks with PAUSE were discussed in the following presentation but more work is needed to address all issues:
  - "Requirements Discussion of Link Level-Flow Control for Next Generation Ethernet" by Gusat et al, January '07
    - au-ZRL-Ethernet-LL-FC-requirements-r03

Discovery and Capability Exchange Protocol

- For the enhanced Ethernet, a mechanism is needed to discover the boundary of the enhanced Ethernet components and exchange capabilities
  - Determine capabilities:
    - priority class (such as bandwidth allocation),
    - congestion management support (optional),
    - per priority PAUSE support, etc.
  - Useful, but not essential, for storage protocols layered over TCP/IP such as iSCSI/iSER
    - Can always fall back to legacy Ethernet behavior
  - Critical for storage protocols directly layered over Ethernet such as Fibre Channel over Ethernet
    - Packet loss due to congestion can severely impact throughput and performance
- We need to start developing a list of requirements/objectives for the discovery and capability exchange protocol
- Previous discussion on this topic leverages LLDP
  - "CM Capability Exchange and Discovery" by Wadekar, January '07
    - au-wadekar-cm-discovery-protocol-needs-012407-v3



Summary

- Work on congestion management in 802.1Qau is a good first step
  - But not sufficient for the enhanced Ethernet to become the converged fabric in the datacenter
- We need to start developing a list of requirements/objectives on these other aspects:
  - Priority processing and packet scheduling
  - Per priority PAUSE
  - Discovery and capability exchange protocol

Straw Poll #1

- The CM task group should draft a PAR, 5 criteria and objectives for "transmission selection" for 802.1Q bridges to provide priority grouping and per-group traffic class allocation, for review by IEEE 802.1 at the July plenary
- Results:
  - Yes:
  - No:
  - Abstain:

Straw Poll #2

- I intend to actively contribute to the development of a PAR, 5 criteria and objectives for the "priority grouping" work in 802.1Q spec
- Results:
  - Yes:
  - No:
  - Abstain:

Straw Poll #3

- The CM task group should draft a PAR, 5 criteria and objectives for granular (per priority) link level flow control for 802.1Q bridges for review by IEEE 802.1 at the July plenary
- Results:
  - Yes:
  - No:
  - Abstain:

Straw Poll #4

- I intend to actively contribute to development of a PAR, 5 criteria and objectives for the "per priority link flow control" work in 802.1Q spec
- Results:
  - Yes:
  - No:
  - Abstain: